A new style of news reporting: Wikileaks and data-driven journalism

Baack, Stefan

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

A new Style of News Reporting: *Wikileaks* and Data-driven Journalism

Stefan Baack, M.A. - s.baack@rug.nl; Centre for Media and Journalism Studies, University of Groningen

Published in: *Cyborg Subjects: Discourses on Digital Culture*, edited by Bonni Rambatan and Jacob Johanssen, 113–122. Shoestring Anthologies. [S.l.]: CreateSpace Independent Publishing Platform, 2013

First published online 2011 at Cyborg Subjects.

Abstract

The coverage of *Wikileaks*' huge amounts of leaked data was a challenge for newspapers – they had to figure out how to get stories out of extensive and complex data sets and how to present their findings to readers. The result significantly differs from traditional news reporting; including illustrations, interactive web applications and reading instructions to make the material accessible. This style of news reporting is called data-driven journalism. The international interest in the leaks combined with collaborative work between newspapers from different countries made it a new trend in current journalism. A key lesson from working with this kind of material is that data collection is essential for the effectiveness of the used techniques. If journalists would adapt this insight to their own, internal data collection process, this form of news reporting could be used on a large scale and be much more common. The coverage of *Wikileaks*' might give a glimpse of how journalism will look like in the future.

A new Style of News Reporting. Wikileaks and Data-driven Journalism

Newspapers are still struggling with the changing media environment that is undermining their traditional business model and are unsure how to make profits online (Freedman 2010). With growing commercialization, journalists tend to use new technology foremost to speed up the news production process rather than experimenting with the new possibilities or enhancing quality (Phillips 2010). However, the collaboration with *Wikileaks* challenged traditional newspapers and forced them to think about new ways of finding and telling stories. They had to work with large and extensive data sets. To take an example, the Afghanistan War Logs consisted of about 92,000 documents written in a military jargon (Rogers 2011). The obvious problem is accessibility – both for journalists who want to get a story out of the material and for readers who want to take a closer look at it. Letting journalists go through everything individually would be too time consuming and writing about the findings in a traditional manner seemed insufficient for the coverage. Especially *The Guardian* and *New York Times* realized that early on. Tools were used to go through the data and to create visualizations and interactive web application which made the material accessible for readers. This form of news reporting is called data-driven journalism – and *Wikileaks* contributed to its development as a trend.

Data-driven Journalism

Scholars and professionals started to discuss data-driven journalism very recently. In April 2010, the *European Journalism Center* and the *University of Amsterdam* initiated the one day event Data-driven journalism: What is there to learn? to define it and discuss possible implications. At this event, Lorenz defined data-driven journalism as “a workflow, where data is the basis for analysis, visualization and – most important – storytelling” (2010: 10). Due to the storytelling aspect, the end product is more than just a visualization of data – it is also contextualizing and highlighting of important aspects.

Bradshaw (2010) explains this data-driven workflow in more detail and distinguishes four steps: finding the data (1), interrogating data (2), visualizing data (3) and mashing data (4). Finding can involve having expert knowledge, good contacts or technical skills to gather data. The interrogation requires a good understanding of the used jargon and wider context of the data. Visualization and mashing can involve the work of designers and/or free tools. An example is *IBM*'s ManyEyes, where users can easily upload and visualize data for free. As Bradhsaw points out, these four steps require teamwork: "The reality is that almost no one is doing all of that" (2010). At the end of this workflow, raw data should be accessible for readers. Lorenz describes it as a process of refinement, raw data is transformed into something meaningful: "As a result the value to the public grows, especially when complex facts are boiled down into a clear story that people can easily understand and remember" (Lorenz 2010: 12).

Data-driven journalism is not something completely new. As Rogers (2010a) shows, it can be considered to be quite old instead. He describes Florence Nightingale as one of the first data-journalists in the 19th century who already worked with visual presentations of information to tell stories. What really *is* new, however, is the media environment journalists are working in. Especially these four aspects indicating a growing importance of data-driven journalism:

- The sheer amount of publicly relevant data available online. Especially in the United States and Britain, huge data sets are available in connection with the open government initiative. The problem here is the same as described above: Having access is not enough without accessibility. To take Britain, most governmental data is released as a simple and static PDF file (Stay 2010). Journalists from *The Guardian* and *New York Times* saw the potential and started to fill this gap by offering interactive tools and illustrations to add public value to the data.
- The existence of free tools to handle this data, like the already mentioned *ManyEyes*.
- The possibility to make the data accessible in an interactive way with web applications.

- Time is precious for journalists, they are always under pressure to get the story out fast (see Phillips 2010). By giving access to the raw data, it is possible to involve people outside the newsroom in the process of news production with crowdsourcing – the collaborative analysis by volunteers. This can save time and resources for researching.

Obviously, data-driven journalism greatly benefits from the possibilities of new media. Its perception as a trend is therefore not surprising.

The role of Wikileaks for Data-driven Journalism

Is *Wikileaks* data-driven journalism in itself? Two contra arguments are that it does not provide visualizations and does not attempt to generate stories out of its materials (only a brief contextualization is given) - both is largely left over to established news media or is considered to be done by 'users' (see Lovink et al. 2010). In regard to the workflow of data-driven journalism, *Wikileaks* is doing the first and second step of collecting and interrogating data without going further. A key aspect, the transformation of raw data into something meaningful to add public value, is not given. To what extent *Wikileaks* can be considered journalistic more generally remains open for debates, but it is not a form of data-driven journalism alone - but surely an important actor in the data-driven workflow nonetheless. From this perspective, *Wikileaks* is a source for data that needs to be 'refined' to add public value.

Wikileaks as a data-source can be called a driving force of data-driven journalism and has contributed to its development as a trend for three main reasons. First and obviously, to analyze and cover its huge amounts of leaked (raw) data, data-driven journalism techniques are essential both for journalists who want to get a story out and present it to their readership and for readers who can access the material through visualizations and reading instructions. The second reason is that the leaks were interesting for an international audience. The released data from the open government initiatives in the United States and Britain were only interesting for national audiences and there was no

need for foreign newspapers to work with it. Connected to this, the third reason is the collaborative work between newspapers from different countries combined with the simultaneous release date of their coverage. The coverage of the Afghanistan War Logs therefore *internationally* demonstrated the advantages data-driven journalism can have. In comparison, not all of *Wikileaks*‘ media partners were able to keep up with *The Guardian* and *New York Times*. In Germany, where the open government movement was (and still is) much weaker, *Der Spiegel* covered the Afghanistan War Logs in a much more ’traditional’ way, using no interactive illustrations at all and focusing on the print version (Krebs 2010). The experience in Britain and the United States to work with huge amounts of data was clearly an advantage for the coverage and made newspapers from other countries aware of the potential. As a result, almost every media partner followed their example and offered visualizations for the second major leak, the Iraq War Logs. As Simon Rogers from *The Guardian* states: “Wikileaks didn’t invent data journalism. But it did give newsrooms a reason to adopt it” (2011).

Using data-driven journalism on *Wikileaks*’ materials: What was there to learn?

To be more concrete about how data-driven journalism was used in connection with *Wikileaks*, lets take a closer look at the Iraq War Logs and the ‘Cablegate’ (focusing on *The Guardian* as an example).

The War Logs contained 391,832 field reports from soldiers. Since each report describes only a single incident, visualizations are extremely helpful to see patterns and get a bigger picture. Two important characteristics made it relatively easy to automatically separate those logs into categories: The standardized format and the use of a dense military jargon, giving meta-data about date, location, type of incident etc. (Matzat 2010). In other words: The data set was largely readable for machines. *The Guardian* concentrated on incidents where someone had died and separated them into cause of death, who were killed (for example civilians or hostile forces), time, location etc.

(Rogers 2011). Then they used *Google Fusion tables* and marked every single death in *Google Maps*. The map was released alongside with key findings from their statistical analysis (Rogers 2010b). This gave an overview of the amount of people killed and further information to contextualize it (for example, most of these people were civilians). In addition, *The Guardian* took all incidents from a single day to create an interactive graphic (Dant et al. 2010). While a timer is running from the first to the last minute of this day, a map shows the location of each incident, gives a description of what happened and counts the total amount of dead people. It also offers a link to the original report of each incident. As Lorenz described, abstract numbers were broken down into something meaningful. By visualizing a single day, you can get a better picture of the atmosphere and violence that shines through the logs. Apart from that, the fact that the material was readable for machines did not only help to create visualizations to present the news and make the material accessible for readers. The automatic separation into categories was used to guide the selection of documents worth reading for the coverage - which can speed up the generating of stories out of the data set.

Compared to the War Logs, visualizations for the ‘Cablegate’ are rare. According to Matzat (2010), this is not only due to the broad geographical reference but mainly to the content of the material. While the War Logs could be categorized and visualized relatively easy due to their clear structure, the diplomatic dispatches (‘cables’) are extensive reports and complex analysis. As Rogers from *The Guardian* points out, their “reporters ended up with the enormous task of actually going through each cable, reading it and seeing what stories were there” (2011). Still, *The Guardian* created a static world map showing how many cables come from which locations and how they are classified. This may be useful to get an overview of the material, but without knowing the actual content of the cables it does not give readers a better access to it. The fact that 1,083 cables have been sent from London to Washington is not interesting without knowing what is written in it. Seeing the problem, *The Guardian* also offers a more ‘context-rich’ interactive map. Users can click on a country and get list of both the original cables from *Wikileaks* and a list of articles covering the content of those cables, which is a very useful

tool to investigate the material. However, only a small amount of cables is available on this map yet, partly due to the material and to the releasing policy of *Wikileaks* (not all cables have been released simultaneously, they continue to be steadily released in stages). For this kind of unstructured material, crowdsourcing or alternative web resources for investigating it is still an advantage of data-driven journalism. There are a couple of crowdsourcing projects or search engines for the cable releases, for example *CableWiki* or *CableSearch* (see an overview here). These resources can form the base for further visualization attempts in the future.

The coverage of the Iraq War Logs and the Cablegate showed that the effectiveness of data-driven journalism techniques is dependent on the material at hand. For structured and machine-readable data, they are very helpful for both showing journalists where to find a story in the material and for readers who can get access through visualizations. For more extensive and unstructured data like the diplomatic cables, visualizations are not as useful and there is no other way than reading everything individually.

First Precursor of a new Journalism?

With more and more publicly relevant data available online and a further development of visualization techniques, data-driven journalism is at least likely to become a more established form of news reporting. However, it is questionable if such data will continue to come from *Wikileaks*. The recent release of the Guantánamo Bay files seems to be "very nearly the final" (Gabbatt 2011) cache of the huge data set the platform supposedly obtained from Bradley Manning. I think such persons who have access to those files and are willing to leak it are far from the norm. Even if *Wikileaks* is this initial spark for a 'leaking culture' (which can be assumed due to the rise of more specialized and local leaking platforms like *Greenleaks*) it is unlikely that leaked data with the same impact and size as the Cablegate or the Iraq War Logs will be common. Apart from that, the future of open government initiatives is unclear as well – especially after the budgets for this project have been cut in the United States (Yau 2011). When newspapers solely rely

on the success of leaks and open government, data-driven journalism may remain a niche form of news reporting.

Therefore, I would argue that the real lesson journalists can learn from the collaboration with *Wikileaks* is shown by Kayser-Bril et al. (2011). They suggest that media organizations should not wait for the release of other data sets and, instead, further embrace the opportunities of data-driven journalism by becoming 'trusted data hubs' themselves. They should not only focus on handling *externally* produced data sets, but also develop and structure their own, *internal* database. Even though Kayser-Bril et al. do not refer to *Wikileaks*, they largely take the experience with its materials into account by stressing that the way data is collected is essential. Basically, all content produced by journalists is already data. What has to be changed is the way this data is collected, making it readable for machines and enable journalists to quickly analyze large and complex data sets and build stories around them. Every event can be broken down by some fundamental information (latitude, longitude etc.), described in a structured manner and linked to other events in a database. As an example of the possibilities, Kayser-Bril et al. mention the crime page of a newspaper. Instead of just giving a list of articles about crime events, it could be transformed into a web application that plots the events over time with the options to sort the data by time, type of crime, location and visualizing it on a map - similar to *The Guardian*'s map for the War Logs.

When newspapers adopt these ideas, data-driven journalism will surely be a more common and established form of news reporting that can come into use regardless of leaks or open government. Journalism could benefit from the new possibilities for finding, telling and presenting stories demonstrated in the coverage of *Wikileaks*' material on a large scale. As Phillips (2010: 100) and Benson (2010: 192) are pointing out, more important than the capabilities of new technology is the way journalists actually use it. Becoming data-hubs could make them aware that they can and should use the new possibilities to improve the quality of news reporting and not only the speed of production. This would be an important step forward - not least initiated due to *Wikileaks*.

References

Benson, Rodney (2010): Futures of the News: International Considerations and Further Reflections. In: Fenton, Natalie (ed.): New Media, Old News. Journalism & Democracy in the Digital Age. London: Sage, P. 187-200.

Bradshaw, Paul (2010): How to be a data journalist. http://www.guardian.co.uk/news/datablog/2010/oct/01/data-journalism-how-to-guide (last accessed 16.04.2011).

Dant, Alastair/Meek, James/Santos, Mariana (2010): Iraq war logs: A day in the life of the war. http://www.guardian.co.uk/world/interactive/2010/aug/13/iraq-war-logs (last accessed 19.04.2011).

Freedman, Des (2010): The Political Economy of the 'New' News Environment. In: Fenton, Natalie (ed.): New Media, Old News. Journalism & Democracy in the Digital Age. London: Sage, P. 35-50.

Gabbatt, Adam (2011): Guantánamo Bay files - live coverage. http://www.guardian.co.uk/world/blog/2011/apr/25/guantanamo-bay-files-live-coverage (last accessed 26.04.2011).

Kayser-Bril, Nicolas/Lorenz, Mirko/McGhee, Geoff (2011): Media Companies must become trusted Data Hubs. http://owni.eu/2011/02/28/media-companies-must-become-trusted-data-hubs-catering-to-the-trust-market/ (last accessed 27.03.2011).

Krebs, Malte (2010): Spon-Chef Rüdiger Ditz zum Blogger-Bashing. "Wir haben nicht so gut ausgesehen". http://meedia.de/nc/details-topstory/article/wir-haben-nicht-so-gut-ausgesehen_100029332.html (last accessed 27.03.2011).

Lorenz, Mirko (2010): Status and Outlook for data-driven journalism. In: European Journalism Center: Data-driven journalism: What is there to learn? A paper on the data-driven journalism roundtable held in Amsterdam on 24 August 2010, P. 8-17. http://mediapusher.eu/datadrivenjournalism/pdf/ddj_paper_final.pdf (last accessed 22.03.2011).

Lovink, Geert/Riemens, Patrice (2010): Twelve theses on WikiLeaks. http://www.eurozine.com/articles/2010-12-07-lovinkriemens-en.html (last accessed 15.04.2011).

Matzat, Lorenz (2010): Wie Wikileaks inzwischen Transparenz versteht. http://blog.zeit.de/open-data/2010/11/29/wikileaks-embassyfiles-transparenz/ (last accessed 17.04.2011).

Phillips, Angela (2010): Old Sources: New Bottles. In: Fenton, Natalie (ed.): New Media, Old News. Journalism & Democracy in the Digital Age. London: Sage, P. 87-101.

Rogers, Simon (2010a): Florence Nightingale, datajournalist: information has always been beautiful. http://www.guardian.co.uk/news/datablog/2010/aug/13/florence-nightingale-graphics (last accessed 22.03.2011).

Rogers, Simon (2010b): Wikileaks Iraq: data journalism maps every death. http://www.guardian.co.uk/news/datablog/2010/oct/23/wikileaks-iraq-data-journalism (last accessed 25.03.2011).

Rogers, Simon (2011): Wikileaks data journalism: how we handled the data. http://www.guardian.co.uk/news/datablog/2011/jan/31/wikileaks-data-journalism (last accessed 16.04.2011).

Stay, Jonathan (2010): How The Guardian is pioneering data journalism with free tools. http://www.niemanlab.org/2010/08/how-the-guardian-is-pioneering-data-journalism-with-free-tools/ (last accessed 22.03.2011).

Yau, Nathan (2011): Data.gov in crisis: the open data movement is bigger than just a site. http://www.guardian.co.uk/news/datablog/2011/apr/05/data-gov-crisis-obama (last accessed 26.04.2011).